



**André  
Ribeiro  
Alves**

**Relatório de Estágio Curricular em Gestão de Dados  
Clínicos**

**Curricular training Report in Clinical Data  
Management**





**André  
Ribeiro  
Alves**

## **Relatório de Estágio Curricular em Gestão de Dados Clínicos**

### **Curricular training Report in Clinical Data Management**

Relatório apresentado à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Mestre em Biomedicina Farmacêutica, realizado sob a orientação científica do Doutor Rui Melo, Director executivo da Eurotrials e da Doutora Maria Amparo Ferreira Faustino, Professora Auxiliar do Departamento de Química da Universidade de Aveiro



Dedico este trabalho aos meus pais, irmã, restante família e namorada pelo incansável apoio.



## **o júri**

Presidente	Doutor Bruno Miguel Alves Fernandes do Gago, Professor Auxiliar Convidado, Universidade de Aveiro
Vogal - Arguente Principal	Doutor Francisco Luís Maia Mamede Pimentel, Professor Associado Convidado C/ Agregação, Universidade de Aveiro
Vogal - Orientador	Professora Doutora Maria do Amparo Ferreira Faustino, Professora Auxiliar, Universidade de Aveiro





## **palavras-chave**

Eurotrials, Estágio, Ensaio Clínico, Gestão de Dados Clínicos, Padronização de Dados

## **resumo**

Este relatório descreve as actividades desenvolvidas no contexto do estágio de 9 meses realizado na Unidade de Gestão de dados da Eurotrials com início em Setembro de 2011 e fim em Maio de 2012.

A Eurotrials é uma empresa de consultoria científica que presta serviços à indústria farmacêutica e biotecnologia, nomeadamente na condução de ensaios clínicos.

No processo do desenvolvimento de um novo medicamento os ensaios clínicos são a ferramenta mais importante de forma a verificar a segurança e eficácia da substância.

A gestão de dados clínicos tem um papel muito importante na condução de ensaios clínicos e tem como objectivo gerar dados de grande qualidade e robustos para que possam ser analisados. A equipa de gestão de dados participa em actividades que vão desde o planeamento do estudo até à sua conclusão.

As principais actividades exercidas no âmbito da gestão de dados foram o desenho do caderno de recolha de dados, bem como a criação da base de dados, gestão de discrepâncias e padronização de dados.



**keywords**

Eurotrials, Curricular training, Clinical Trial, Clinical Data Management, Data Standards

**abstract**

This report describes the activities undertaken in the context of a Curricular training with a duration of 9 months in the Data Management Unit at Eurotrials starting in September 2011 and end in May 2012.

Eurotrials is a contract research organization that provides services to the pharmaceutical and biotechnology industries, namely clinical trial conduction.

In the drug development process, clinical trials are the most important tool to verify if the drug is secure and effective.

The field of clinical data management has a very important role in clinical trials conduction and aims to generate high-quality and reliable data so that it can be analyzed. The data management team is engaged in activities ranging from the design of the study until their completion.

The main activates performed regarding clinical data management were the design of the case report form, database design, discrepancies management and data standardization.



## Table of Contents

List of abbreviations .....	3
1. Introduction.....	5
1.1 Training Objectives .....	5
1.1.1 Primary objectives .....	5
1.1.2 Secondary objectives .....	5
1.2 Vision of the Host Institution.....	6
2. State of the Art .....	8
2.1 Importance of Clinical Data Management in the Outcome of Drug Development .....	8
2.2 Clinical Trials – An Overview.....	8
2.3 Development Phases and Types of Study.....	10
2.3.1 Phase I (Human Pharmacology).....	10
2.3.2 Phase II (Therapeutic Exploratory) .....	10
2.3.3 Phase III (Therapeutic Confirmatory) .....	11
2.3.4 Phase IV (Variety of Studies: - Therapeutic Use) .....	11
2.3.5 Methods to Minimize or Assess Bias and improve data quality.....	11
2.4 Contract Research Organization Market .....	11
3. Clinical Data Management Activities.....	13
3.1 Data Management Plan .....	13
3.2 Clinical Data Management Systems .....	13
3.2.1 Audit trail.....	14
3.3 Case Report Form Design .....	14
3.3.1 General organization of data collection forms.....	15
3.3.2 Content, Presentation and Methodology.....	16
3.3.3 Wording.....	18
3.3.4 Minimizing redundancy.....	19
3.3.5 CRF completion Guidelines.....	19
3.4 Database Design and Specification.....	19
3.4.1 Calculated or Derived data .....	20
3.4.2 Database structure .....	20
3.4.3 Database Specification .....	22
3.4.4 Database Validation.....	22
3.5 Data capture and Data entry .....	23
3.5.1 Workflow .....	23
3.5.2 Data receipt .....	23
3.5.3 Data entry.....	23
3.5.4 Paper Case Report Forms considerations.....	25
3.5.5 Data entry screen design.....	25
3.5.6 EDC considerations.....	26
3.5.7 Modifying data .....	26
3.6 Electronic Data Capture.....	26
3.6.1 Differences between Electronic Data Capture and Paper-Based studies.....	27
3.6.3 Advantages and disadvantages of Electronic Data Capture .....	28
3.7 Data Validation Procedures .....	29
3.7.1 Defining and Implementing Edit Checks.....	29
3.7.2 Managing Discrepancies.....	30
3.7.3 Turning discrepancies into queries.....	31
3.7.4 Queries Resolution .....	31
3.8 Database Quality Control .....	32
3.8.2 Data Error Categorization .....	32
3.8.2 Data sample definition .....	33
3.8.4 Error rate calculation.....	33
3.8.4 Corrective actions.....	33

## Curricular training Report in Clinical Data Management

3.8.5 Documentation .....	33
3.9 Final Database .....	34
3.9.1 Database Closure Procedures .....	34
3.9.2 Errors found after database closure .....	34
3.9.3 Randomization code break .....	35
4. Data Standards and CDISC Implementation.....	36
4.1 Data Management Standards in Clinical Research .....	36
4.1.1 Barriers to Standards Implementation.....	36
4.1.2 Purpose and Benefits of Standardization.....	36
4.1.3 Data Standards categories .....	37
4.2 Clinical Data Interchange Standards Consortium.....	38
4.3 Study Data Tabulation Model Overview .....	39
4.4 CDISC Standards Implementation .....	44
5. Curricular Training Timeline.....	46
6. Conclusion .....	51
7. References.....	52

## Figures and Tables

Figure 1- Eurotrials Organogram (2) .....	6
Figure 2- Data Management Unit Organogram (2).....	7
Figure 3 - Clinical Development (5).....	9
Figure 4- Comparison of R&D investment and global output of new molecular entities (NME).(7) .....	12
Figure 5 - Curricular Training Timeline.....	46
Table 1 - Data stored in a non-normalized or short-fat table .....	21
Table 2- Data stored in a normalized or tall-skinny table .....	21
Table 3 - Data stored in a hyper-normalized table .....	21
Table 4 - Acceptable error rates.....	33
Table 5 - SDTM Domains(22, 23).....	40
Table 6 - Define.xml Table of Contents(24).....	42
Table 7 - Define.xml Data Definition Table(24) .....	43
Table 8 - Maximum size of variables and data(25) .....	43
Table 9 - Studies Description.....	48

### List of abbreviations

AE	Adverse Events
CDASH	Clinical Data Acquisition Standards Harmonization
CDISC	Clinical Data Interchange Standards Consortium
CDM	Clinical Data Management
CDMS	Clinical Data Management System
CRA	Clinical Research Associate
CRF	Case Report Form
CRO	Clinical Research Organization
DCF	Data Clarification Form
DDD	Data Definition Documentation
DDT	Data Definition Tables
DEO	Data Entry Operator
DMP	Data Management Plan
DVP	Data Validation Plan
e-CRF	Electronic Case Report Form
EDC	Electronic Data Capture
EMA	European Medicines Agency
FDA	Food and Drug Administration
GCP	Good Clinical Practices
GUI	Graphical User Interface
ICH	International Conference on Harmonization
IS	International Standard

## Curricular training Report in Clinical Data Management

NME	New Molecular Entities
QC	Quality Control
SAS	Statistical Analysis System
SEC	Self-Evident Corrections
SDTM	Study Data Tabulation Model
SDV	Source Document Verification
SME	Small Medium Enterprise
TOC	Table of Contents
WHO	World Health Organization



## 1. Introduction

This report reflects the work developed during 9 months at Data Management Unit of Eurotrials. This curricular training occurred during the second year of the Master's degree in Pharmaceutical Biomedicine.

### 1.1 Training Objectives

#### *1.1.1 Primary objectives:*

- To Improve and the knowledge acquired during the degree in biomedical sciences and during the Master's degree in Pharmaceutical Biomedicine;
- To obtain specific working skills and techniques;
- To understand the work environment of an organization;
- Establish an integrated working approach, with the constant communication between clinical trials unit and data management unit;
- To establish working contacts network.

#### *1.1.2 Secondary objectives:*

- Identify and understand the importance of data management in clinical trials context;
- Develop knowledge and expertise in the data management activities performed throughout clinical research including:
  - Understand how clinical data is analyzed and transformed in comprehensive information for statistical analysis;
  - Contribute to the implementation of CDISC guidelines in clinical data management at Eurotrials;
  - Develop data management plans;
  - Construct annotated CRF, database dictionary and manual;
  - Develop data validation plans;
  - Design, programming and validation of databases;
  - Data entry;
  - Detection of inconsistencies: emission, tracking, receipt and entry of queries;
  - Perform Final quality control for database lock;
  - Export information to database in SAS;
  - Elaborate Tables and listings;

### 1.2 Vision of the Host Institution

Eurotrials Scientific Consultants was established in 1995, it is a CRO specialized in clinical research and scientific consultancy. It is a privately owned company founded by members of academia, medical community and pharmaceutical industry.(1)

At Portugal, Eurotrials is certified and recognized by several institutions and organizations, among them: (1)

- ISO 9001 quality certification.
- Innovation Small Medium Enterprise (SME) network COTEC since May 2007.
- Leading SME: In September 2007, IAPMEI recognized Eurotrials as a leading SME.

The main services provided by Eurotrials are:(2)

- Research Methodology Consultancy;
- Feasibility Analysis;
- Trial Design, Protocol and Case Report Form (CRF) Development and Implementation and Monitoring of Clinical Trials;
- Epidemiology and Late Phase Research Studies and Health Economic Studies;
- Data Management and Statistics;
- Regulatory Affairs and Pharmacovigilance;
- Research and Development Consultancy Services;
- Medical Writing;
- Quality Assurance;
- Teaching and Training Activities

The several activities performed at Eurotrials are represented in the following organizational organogram.

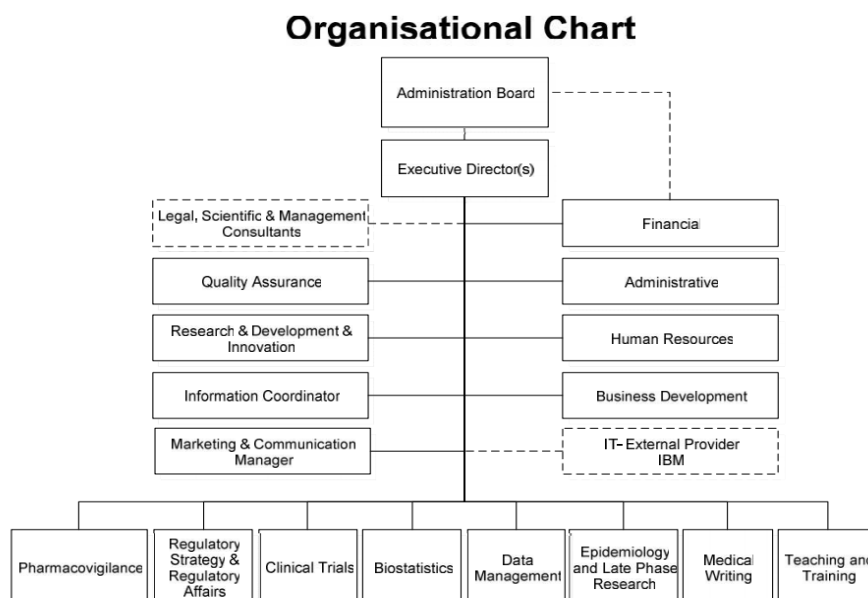
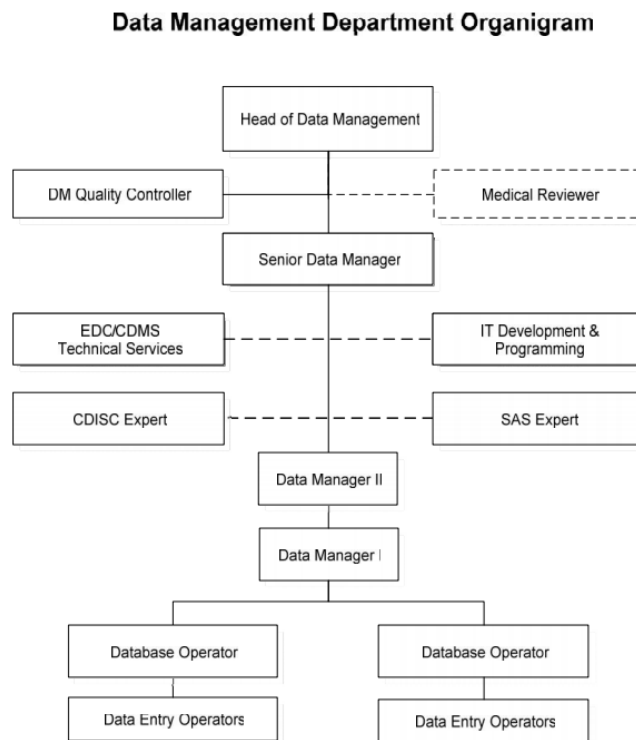


Figure 1- Eurotrials Organogram (2)

## Curricular training Report in Clinical Data Management

The following organogram represents the structure of Data Management Unit at Eurotrials.



**Figure 2- Data Management Unit Organogram (2)**

The data management team at Eurotrials uses innovative technology in electronic data collection, which leads to an effective analysis of the data. The Eurotrials' data management system has been validated and is in strict compliance with regulatory authority guidance's.(2)

## **2. State of the Art**

### **2.1 Importance of Clinical Data Management in the Outcome of Drug Development**

Clinical trials are the most important tool in drug development and those are intent to find the answers to the questions proposed at the beginning of the drug development. Therefore it is important to develop the appropriate studies and apply the right tools in analyzing the data resultant from clinical trials. This data is generated with the goal of proving or disproving a hypothesis. The quality of the data that is generated in clinical trials is very important for the outcome of the drug development and therefore it must be carefully collected, managed and reported. Is in this point that clinical data management is important to collect, cleaning, and manage the subject data in compliance with regulatory standards. Its primary objective is to generate high-quality data by keeping the number of errors and missing data as low as possible, and thus prepare the data for the subsequent statistical analysis.(3)

For this purpose, it is important to establish a consistent and updated data management plan in each drug development, since it is a relevant and important part of a clinical trial

It is also important to meet all the regulatory and guidance documents applied to data quality and when managing clinical data.(3)

When from the process of data management results high-quality data, the statistical analysis is facilitated and the results are more accurate and suitable of the clinical reality. Therefore high-quality data means minimal or no misses, but more important it should possess only a certain level of variation between the captured data and the real data, that would not affect the conclusion of the study on statistical analysis.(3)

### **2.2 Clinical Trials – An Overview**

The definition of Clinical Trial, according to the World Health Organization (WHO) is the following: “a clinical trial is any research study that prospectively assigns human participants or groups of humans to one or more health-related interventions to evaluate the effects on health outcomes. Interventions include but are not restricted to drugs, cells and other biological products, surgical procedures, radiological procedures, devices, behavioral treatments, process-of-care changes, preventive care, etc.”

Clinical trials are required for five purposes:(4)

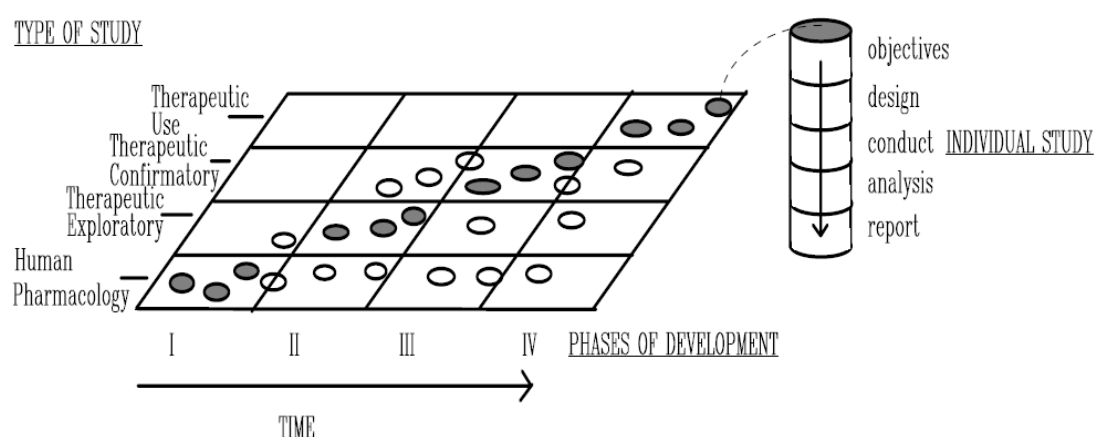
- Guide the drug development and move a drug through it.
- Gain marketing authorization
- Investigate a certain property of the drug
- Select one drug rather than another
- Guide treatment of individual patients

## Curricular training Report in Clinical Data Management

Clinical trials can be classified according to when the study occurs during clinical development or according to their purpose.(5) The results of prior studies should influence and shape the plan of later studies. Generally study results will lead to a modification of the development strategy.

Clinical trial development is based in phases, although it is not the most correct description, since it may lead to the erroneous notion that trials occur in a sequential mode and that each type of study only occurs in one phase and that studies from different phases doesn't occur at the same time. One type of trial may occur in several phases. Consequently it is important to comprehend that the temporal phases do not imply a fixed sequence of studies and for most of drugs the typical sequence will not be appropriate or necessary.(5)

The best way to classify a clinical trial is through objectives, as outlined in the figure 1.



**Figure 3 - Clinical Development (5)**

This graph represents the clinical development, with the shaded circles showing the types of studies most usually conducted in a certain phase of development, the open circles show certain types of study that may be conducted in that phase but are less usual.(5)

The first studies that should be performed in a clinical development should identify characteristics of the investigational medicine to plan an appropriate development based on this profile.(5)

The small and early clinical trials in healthy volunteers provide important information to develop appropriate bigger and more definitive clinical trials.

The first human trials aim to establish a safety profile and tolerability evaluation, as well information needed to choose a suitable dosage range and administration schedule for initial exploratory therapeutic trials.(5)

After the completion and statistical analysis of the exploratory data, confirmatory studies are performed to establish an efficacy profile of the drug in certain populations and evaluate the long-term effects of it. These studies are generally larger and longer and include more diverse patient population.(5)

### 2.3 Development Phases and Types of Study

#### 2.3.1 Phase I (Human Pharmacology)

This phase starts with the initial administration of an investigational new drug into humans.

The most typical studies conducted in this phase are the human pharmacology studies; however they may be present at other point times during the clinical development. The main goals of these studies usually are related with the safety profile and tolerability of the drug; therefore they don't have therapeutic objectives and, as referred previously, may be conducted in healthy volunteers. However drugs that are highly toxic are studied in patients to minimize the consequences to healthy volunteers.(5)

Main aspects analyzed during phase I studies:(5)

- **Estimation of Initial Safety and Tolerability:** the first studies in humans are usually performed to analyze the safety of a medicine and to determine the nature of adverse reactions that can be expected. These studies may be performed with a single or multiple dose administration.
- **Pharmacokinetics:** the pharmacokinetic profile of drug depends on various factors such as the formulation and type of administration. Therefore the preliminary characterization of pharmacokinetics is an important objective of Phase I. The results of these studies will generate information about the clearance of the drug and allow the anticipation of possible accumulation of parent drug or metabolites and potential drug-drug interactions. Nonetheless, it is important to retain that the absorption, distribution, metabolism, and excretion characteristics of a drug continue during all clinical development.
- **Assessment of Pharmacodynamics:** early pharmacodynamic studies in patients may provide information about the activity and potential efficacy of the drug and the dosage regime necessary to promote a positive therapeutic response and the dosage for the next.
- **Early Measurement of Drug Activity:** these studies are usually carried in later phases, however when drug activity may be measured with a short duration of drug exposure in patients, these studies are a good tool to decide to continue or not with the clinical development.

#### 2.3.2 Phase II (Therapeutic Exploratory)

This phase has the main goal to explore therapeutic efficacy in patients and its principal objective is to determine the dose(s) and regimen for Phase III trials. Additional objectives may include evaluation of potential study endpoints, therapeutic regimens for further studies in later phases.(5)

The designs of these studies include concurrent controls and comparison with baseline status for the initial trials in this phase; however when the aim is evaluate the efficacy of the drug and its safety for a particular indication, then they are randomized and concurrently controlled in a group of patients who are selected by very restricted criteria, which makes the study population homogeneous.(5)

Dose regimes design: Early studies in this phase use dose escalation designs to estimate the dose response profile. The dose response profile is confirmed in later studies, though parallel dose-response designs. (5)

### *2.3.3 Phase III (Therapeutic Confirmatory)*

These studies are larger and their duration is extended for years, their major goal is to demonstrate, or confirm therapeutic benefit, in other words, to confirm if the data obtained from Phase II is correct and that the drug is safe and effective for use in the intended indication and population. These are the studies conducted to provide an adequate basis that the drug is safe and effective.(5)

### *2.3.4 Phase IV (Variety of Studies: - Therapeutic Use)*

Phase IV begins after drug's approval. These studies are not consider necessary for approval but are often important for improving and optimizing drug's use. This kind of studies is related to the approved indication.(5)

### *2.3.5 Methods to Minimize or Assess Bias and improve data quality*

The quality of the information collected during clinical trials may be enhanced through some methods, described below.

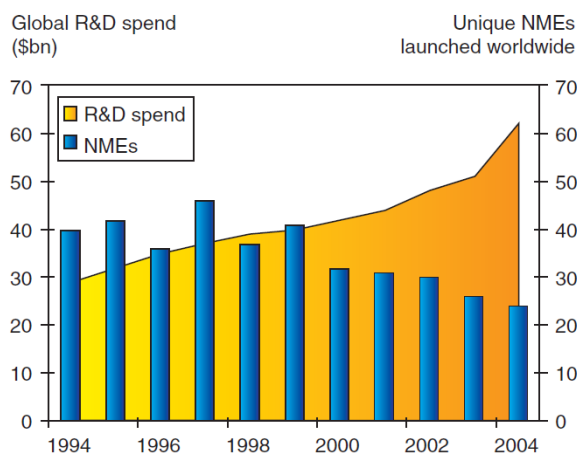
- **Randomization:** in a control trial, this is the preferred means of assuring comparability of test groups avoiding or minimizing the possibility of selection bias.(5)
- **Blinding:** is an important method used to minimize the risk of biased study outcomes. Blinded studies are controlled trials, where an active control or placebo may be used. When only the participant doesn't known each treatment he/she is on, it is called a single blind, when the investigator and sponsor staff who are involved in the treatment aren't aware of what treatment each patient is receiving, it is called a double-blind.(5)

## **2.4 Contract Research Organization Market**

In the past few years, the pharmaceutical industry has been facing productivity crisis. The investment in R&D has been rising, however the productivity and successful development of novel drugs is slowing.(6)

## Curricular training Report in Clinical Data Management

On an effort to reduce the costs with R&D, the pharmaceutical industry will increase the use of third parties, such as Contract Research Organizations (CRO).



**Figure 4- Comparison of R&D investment and global output of new molecular entities (NME).(7)**

CRO is an organization that provides support to the pharmaceutical and biotechnological industries in their clinical activities in an outsourced form. CROs have become a major partner of pharmaceutical industry in clinical trial activities, providing substantial global capacity to drug developers.

In 2008 the total CRO market was estimated to be 20 billion dollars and it was expected to grow at an annual rate of 8.55% to reach 35 billion dollars through 2015. This is due to the fact that now, more than ever, pharmaceutical industry needs to implement approaches that allow them to save and capitalize financial resources. The conduct of clinical trials are a good example of how a CRO allows a company to save money, because CROs complete a study up to 30% faster than those that are conducted in-house by pharmaceutical companies.(7)

Thus CROs and pharmaceutical are becoming major associates with and integration of the strategic partnerships to gain a competitive edge in the global business environment.(7)



### **3. Clinical Data Management Activities**

#### **3.1 Data Management Plan**

The data management plan document (DMP) is a road map to handle the data usually placed before the data collection which serve as the authoritative resource, documenting data management practices and decisions to be followed in the trial.(8, 9)

The procedures stated in the DMP needs to comply with all applicable regulatory guidelines, e.g. European Medicines Agency (EMA), Food and Drug Administration (FDA), International Conference on Harmonization (ICH), Good Clinical Practices (GCP), national laws as well with Eurotrials standard operating procedures. (8)This document includes:

- The description of the data management staff
- Functions and authorized tasks of each member of data management staff
- Tasks to be performed according to the financial contract and the study protocol
- Primary and secondary trial objectives as defined by the study protocol
- The software to be used for the clinical data management
- Rules and procedures for database design and implementation
- Procedures, rules and conventions for data entry

The DMP is an auditable document often asked for by client or regulatory inspectors, therefore the DMP is written professionally and with special consideration. During an audit, the inspectors may seek to verify the degree to which the CDM staff adhere to the processes described in the DMP.(8)

#### **3.2 Clinical Data Management Systems**

Clinical data management systems (CDMS) are large and sophisticated applications that are available as a tool for clinical data management activities. Also, these software tools are expensive and need sophisticated information technology structure to properly function. In multicentre trials, a CDMS has become essential to handle the huge amount of data.(9)

CDMS's are designed especially to support clinical data management staff to perform tasks over multiple and simultaneous studies. The CDMS's have an underlying database that is used to store the data collected in clinical trials. On top of the database there is a graphical user interface (GUI) screen, checklist or form

where CDM staff performs their activities, the CDMS's will translate these activities into the tables defined in the database.(9)

CDM activities at Eurotrials are performed with these software tools, which support a vast range of data management tasks including:(9)

- Database design
- Entry screen creation
- Data entry
- Data cleaning
- Data validation through edit checks (automatic and manual)
- Discrepancy management and query resolution
- Locking of studies
- Extraction of data for reporting and analysis

### *3.2.1 Audit trail*

Audit trail is a secure, computer generated, time-stamped electronic record that allows reconstruction of the course of events relating to the creation, modification, and deletion of an electronic record. Audit trail of CDM activities is a complex and one of the most important features that CDMS can offer. When using CDMS multiple user IDs can be created with access limitation to the data, audit trail ensure that each user can only perform the activities assigned to them. When clinical data management procedures require to change the original data, the audit trail will keep the track of the old value,, the user ID that made the change, the time of change, and the reason to justify the change. Thus, during an audit, auditors can verify if the changes to the original data were allowed and accurate.(8, 9)

## **3.3 Case Report Form Design**

Although the study protocol is considered the most important document used during a clinical trial, study case report forms (CRFs) are of vital importance too. Protocol defines the objectives of the study, usually regarding efficacy and safety of a specific Investigational Medicinal Product, being the level of emphasis dependent on the phase of the clinical trial.(10)

The International Conference on Harmonisation's Guidance for Industry: E6 Good Clinical Practice defines the term CRF as, "A printed, optical, or electronic document designed to record all of the protocol-required information to be reported to the sponsor on each trial subject."(11)

The quality of study data is heavily dependent on the quality of the tool used to collect data. CRF's are the most common tool for data collection; therefore special attention must be given to ensuring that data

specified in the protocol are collected accurately and consistently. Also, a well-structured CRF simplifies database design, data validation processes and statistical analysis.(8)

The construction of a CRF receives numerous inputs from sources and departments, among them are data management unit, statistics unit, clinical monitoring and regulatory staff, that will ensure that the data to be collected with CRF meet the needs of the study from all pertinent perspectives.(9)

Whatever medium is used for the CRF, paper or electronic, the CRF can only be as good as protocol and can't compensate its inadequacies. Reviewing the protocol is essential since it provides an overview of the clinical trial and an assessment of possible impacts in CRF design.(10)

### *3.3.1 General organization of data collection forms*

At Eurotrials in the procedure of designing a CRF, the first decision is related with the structure (organization) to be used; forms could be organized on the basis of separate clinic visits or evaluations, with all forms being used to a specific visit (*Visit 1*) separated from the forms of another visit (*Visit 2*). An alternative approach is to base the grouping on integration of similar or identical forms (e.g. forms for all vital signs are placed together to be followed by all laboratory forms, which in turn are followed by all forms for concomitant medication). By this way, different copies of the same form (for any parameter) to be filled in different visits are placed together. An alternative approach, is to combine the previous described approaches and have some forms arranged in a visit-by-visit grouping and other forms integrated by type of examination, test, of form.(12)

#### **Visit-by-visit format**

The advantages of using a Visit-by-Visit format for data collection forms is that is easier for investigators to complete them, especially if the data is recorded on the CRF at the time of patient's visit. One disadvantage, however, of this approach is the difficulty to observe trends in data. For example, a gradually changing efficiency parameter may be overlooked as values of each visit are in different sections of the CRF and data may not so easily compared if all values were on the same contiguous pages.(12)

#### **Combining similar pages**

Contrarily to Visit-by-Visit format, the advantage of combining of similar pages in the same section of CRF is that data monitoring and trend detection are more easily achieved. However, the major problem with this approach is the CRF completion that is more difficult, since the necessary forms that must be completed for a specific visit are in different sections of the CRF.(12)

#### **Hybrid approach**

In certain occasions, the combining of the previous formats may present an advantage. It may be considered important to place specific parameters closely, but the rest of the parameters are formatted by

the Visit-by-Visit approach. At Eurotrials this approach is generally used. For example, the concomitant medication, adverse events, laboratory parameters sections are designed following the combining approach, the measurement of vital signs, administration of treatment or even drug accountability are designed considering the Visit-by-visit approach with each of this three last forms being repeated for each visit that composes the clinical study.(12)

### *3.3.2 Content, Presentation and Methodology*

CRF completion is prone to human error. Improving the clarity and ease of a CRF will improve the quality of data collected in this document.(8) According to Wright and Haybittle there are three main factors which contribute to ensure that a CRF is easily understood and used(13):

- Content – do you need to collect it?
- Presentation – are you asking the question correctly?
- Methodology – what design alternatives are available to avoid/minimise problems that users have?

#### **Content**

All data collected in CRFs must be correctly attributable to a subject. Each page or section of CRF that can be separated or viewed separately must contain enough identifiers to uniquely identify the data contained in the page or section.(8)

The protocol identifies the data to be collected during the trial to achieve the study objectives and meet regulatory requirements. Some studies will legitimately omit a few of the following information, but the majority will be included:(10)

- Date, phase and identification of the trial.
- Identification of the subject
- Age, sex, height, weight and ethnic group of the Subject
- Particular characteristics of the subject (smoking habits, dietary status, fertility/pregnancy status, previous treatments)
- Adherence with inclusion and exclusion criteria
- Medical History
- Dose, dosage schedule, administration of medical product, compliance record.
- Concomitant use of other medicines and non-medicinal interventions/therapy
- Recording of efficacy and or safety parameters met, date and time
- Reasons for withdrawal
- Recording adverse events, description, seriousness, severity, duration, intensity, consequence (outcome) and measures taken.

## Curricular training Report in Clinical Data Management

### Presentation

It is important to anticipate the type of answer that is intended to be given. The types of responses are (10):

- Open – including text, number, alpha numeric
- Closed – including binary and multiple choice
- Combination – of the open and closed response
- Analogue scales – alternative rating response

Open responses are used when the answer cannot be predicted and it is provided space for the written response.(10)

Dates and times are usually of open numerical response. Within data management units that conduct multinational studies, Date Fields are commonly error prone due to different formats: “DD-MM-YY”; “DD-MM-YYYY”; “MM-DD-YY”; “YY-MM-DD”. To avoid this type of errors the date fields must be standardised, usually to the following date format: “DD-MMM-YYYY”. Also, attention is required to the collection of clock-time when time fields are required, they should be recorded using the “24-hour” format and not using the “a.m.” or “p.m.” format.(10)

When the fields to be recorded are identifiers or measurements such as pressures, rates, length character separators can be useful by anticipating their magnitude (e.g. weight measurement |\_|\_|, |\_| Kg).(10)

Usually open character field relates to comment fields to collect the reasons for treatment interruption, patient withdrawal, patient death or adverse events. The problem with open fields is the difficulty to transpose the content in comment field in a way that it can be meaningful analysed.(10)

Closed responses are used when the answer is predicted and a list of options can be provided. The advantages of this kind of response are: the clarification of the question meaning the provided list of answers and it is simpler to make the choice; it also allows the answer to be automatically compatible for computer analysis, and forms of electronic data capture.(10)

In paper trials the selection method is important, the printed list can be annotated to show the correct choice by circling, underlining, ticking or checking boxes or deleting the incorrect choice. Wright and Haybittle discover that ticking and checking is the best form filler because it is the fastest form filler and the easiest read by data entry.(13) When using eCRF, combination boxes (*combo boxes*) and check lists are the most common closed response types. Frequently the closed responses are coded before stored at the database for easier data entry and analysis. The codification of a closed response is done by assigning an alpha numerical character to all possible options for a specific answer. The options of closed responses and the assigned characters are called *codelists*.(10)

Combination response extends the range of the closed format by the addition of an open format. The most common example is “Other, specify” which is the last item in a multiple choice list and is used when all

possible options are not known or not included in the list of choices. “Other, specify” comes at last in order to ensure that all the anticipated options are considered first, that is by process of elimination.(10)

Analogue scales are an alternative rating response. Visual analogue scales (usually horizontal lines of 100 mm) length) are used to measure an individual’s perception of improvement and feelings. Labels defining the range are put at either end of the line and the subject is asked to mark the line at a position which best represents their own situation. The data is interpreted by measuring the mark on the line from one of the ends. However, analogue scales are difficult to set up, monitor and validate, so it is recommended to provide clear instructions for marking the line (if possible with illustrations) and ensure that is exactly 100 mm in length on return form the printers.(10)

### **Methodology**

A CRF flow usually follows the data flow from the perspective of the person filling the form. Furthermore data fields are arranged in a manner that is clear and easy to follow. Data that are logically related are grouped together whenever possible.(10) This group of data, connected between them, is called “Module”. A CRF is composed by different modules including but not limited to: demography, adverse events, concomitant medication and medical history.

While designing the layout of a CRF, CDM staff is very careful with the format and characteristics, since it should be consistently including font size and the use of colour. Attention is also required to the intended use of the form. (8) Therefore, in order to enhance readability and understanding as well the desired use of the form, the choice of layout and design of a CRF should be based in the following characteristics: (10)

- Type sizes: 8-12 point
- Type face: e.g., Times Roman, Helvetica, Ariel, Univers
- Case: mixed, if text is all presented in upper cases the readability can be reduced between 13-20%
- Line length: 40-70 characters; very short and very long lines should be avoided.
- Spacing: when items are less spaced it indicates that are related, more space separate unrelated ones, space between lines should be inferior then between words.

### **3.3.3 Wording**

Besides the layout, attention to wording should also be given when designing a CRF, since the form filler cannot interact with the CRF to obtain clarification and may be unfamiliar with the colloquial speech of the language used. In order to avoid ambiguities, the following aspects are taken into consideration:(10)

- Avoid use of words with more than one meaning.
- Avoid use double negative; it should be used positively worded questions and statements.
- Avoid use of passive voice, instead use active voice as it links individuals with actions allowing a better understanding of CRF by the form filler.

- Avoid extending unnecessarily phrases or statements; when possible replace it with fewer words.
- Avoid leading questions, as subjects may feel intimidated and try to please the investigator.
- When possible, composed questions should be broken into a series of single questions, by using binary questions (Yes/No), orienting the form filler to go to another question if needed.

### *3.3.4 Minimizing redundancy*

Data based on the same measurement should not be collected more than once, because doing so creates unnecessary work for site staff and creates a need to check for consistency between redundant data.(8) (9)(10) In the same manner, collection raw of data is usually better than collecting calculated values, because raw data is easier to verify from source documents and there isn't the risk of potential errors coming from calculated data from different staff and at different time points. An example could be the calculation of *Body Surface Area* or *Body Mass Index*, since it can be easily computed by CDM or statistical staff and recorded on the central database or on the analysis database.(8)

### *3.3.5 CRF completion Guidelines*

To ensure that CRFs are completed correctly, those should include clearly stated instructions on the associated CRF completion guidelines. These guidelines are used to train site staff and help monitors when reviewing data in completed forms. Moreover, CRF completion guidelines encompass instructions regarding methods of correcting and changing data.(8)

Instructions and guidelines on completion should take into account the used data collection method (paper versus EDC) and should be adapted to the individuals who will be filling the CRF since these guidelines can be very different when CRF is to be completed by study subjects or by study staff, generally CRFs are filled by the investigators or study coordinators however patient diaries, questionnaires to assess quality of life as well the use of scales. Also paper-based CRFs have printed completion guidelines, while EDC systems may use on-line help menus despite of printed guidelines. (8)

## **3.4 Database Design and Specification**

Regardless the data capture method, which may be paper, EDC or laboratory instruments, data collected from a clinical trial is stored in some kind of computer system or systems. Databases should be designed to allow complex data to be cleaned, reviewed and reported.(9)

A database is simply considered a set of data that is structured. This could be an Excel spreadsheet, a Microsoft Access, collection of SAS tables or a set of tables built in one of the traditional relational applications such as Oracle.(9)

Independently of the software used to store and manage clinical data, the CDM staff can only design the structure of the database for each clinical trial after the final version of protocol has been defined and the CRF has been drafted.(9)

The main goal for all databases containing information from clinical trials is to store data accurately. Additionally, when designing the database also other aspects are considered, including:(9)

- Clarity, ease and speed of data entry
- Efficient creation of analysis data sets for biostatisticians
- Formats of data transfer files
- Database application software requirements

Commonly, data capture instruments are defined before database implementation. The different types of data capture instruments influence the database design but those don't completely determine it, since data from a given CRF page can be stored in a different number of ways in the database.(9)

### *3.4.1 Calculated or Derived data*

Data from the CRF based on paper or electronic systems are not the only data related with a clinical study. There are internal fields that can be useful and even import to the data processing of data that are calculated from other data that are obtained using mathematical expressions or are derived from other data using text algorithms or other logic.(9) Examples of calculated data include BMI, when weight and height are recorded, or transformation of laboratory units to international standard (IS). Examples of derived data are the auto completion of fields when other specified fields are filled or extracting the site identifier from the log patient identifier.

These fields can be calculated or derived in the central database and others are calculated or derived posteriorly in the analysis database. At Eurotrials, the CDM staff defines where these values should be loaded. If these fields are necessary to discrepancy identification and report generation, the best approach is to include them in the central database. However, if these values are only used during analysis there is no need to create a permanent storage for them on the central database.(9)

### *3.4.2 Database structure*

A very important aspect taken into consideration when designing the database is its normalization.

Database normalization is the process of creating a design that allows efficient access and storage. In some CDM systems, database records are closely related to the implemented CRF page. Therefore, few options for normalization are available to the CRF designer. Other systems, contrarily to the previous, there is a high level of normalization and the designer have no design choices. Finally, there are systems that database designer can choose the level of normalization intended.(9)



## Curricular training Report in Clinical Data Management

The following tables represent a non-normalized and normalized data storage structure:

**Table 1 - Data stored in a non-normalized or short-fat table**

SUBJID	VISIT	BPM_1	BPM_2	BPM_3	TEMP_1	TEMP_2	TEMP_3
01	5	56	60	58	36.8	36.9	37.0

**Table 2- Data stored in a normalized or tall-skinny table**

SUBJID	VISIT	MEASURE	BPM	TEMP
01	5	1	56	36.8
01	5	2	60	36.9
01	5	3	58	37.0

The normalized version of the table has fewer columns and more rows than the non-normalized version. The visual impact of the normalization in the database has led to call to them in CDM jargon as *short-fat* for non-normalized tables and *tall-skinny* for normalized tables.(9)

Both types of structures store the data accurately and allow its retrieval for analysis, but the choice impacts the data management activities. As example, detecting missing records is easier in non-normalized tables but creation of the structures themselves and the related checks are facilitated when using normalized structures. The advantages from normalized-tables come from the fewer, uniqueness and clear names of the fields that compose the table.(9)

Generally, clinical data collected in a tabular form is easily transferable to a normalized table. Adverse event forms, concomitant medication and lesions assessment are examples of data often collected in a tabular way and stored in a normalized way.

In the example given for the normalized form, each column "BPM" and "TEMP" record data is different between them in format and units. However, normalized forms are so flexible that it allows each column to collect data of different types of measurement. This approach is called "hyper-normalization". This "hyper-normalization" structure is similar to the Study Data Tabulation Model structure proposed by CDISC (Clinical Data Interchange Standards Consortium). Laboratory and vital signs are the typical example of this type of normalized form, where one column may give the test name, other, the test result, and another, the units. The "hyper-normalization" of the example given for a normalized form would result in the following table:

**Table 3 - Data stored in a hyper-normalized table**

SUBJID	VISIT	MEASURE	TEST	RESULT	UNITS
01	5	1	HEART RATE	56	BMP
01	5	1	TEMPERATURE	36.8	°C
01	5	2	HEART RATE	58	BMP
01	5	2	TEMPERATURE	36.9	°C
01	5	3	HEART RATE	60	BMP
01	5	3	TEMPERATURE	37.0	°C

### *3.4.3 Database Specification*

The output of the database design process is its specification. The specification for a given study is, at a minimum, an annotated CRF. Additionally to the annotated CRF, data definition documentation (DDD) is frequently required.(9)

Annotated CRF is usually a blank CRF that has written on it, by hand or in electronic way, the names of the fields, or item associated with each CRF field. In different words the annotated CRF maps each field on the CRF with the corresponding variables on the database.(9) The annotated CRF is also clearly marked to show how questions are related or grouped into modules or tables. It is also helpful, to make the annotation of codelists related to closed fields as well as the annotation of any hidden, internal or derived fields associated to each module. Annotated CRF is not only used by database designers. This document will be used along CDM activities by different staff including edit checks, edit check designers, entry screen designer and those searching for inconsistent data on the database.

The DDD is a technical document that provides a list of all modules or tables, the fields contained in each module, the codelists associated with closed fields and its codification. Also, the attributes for all fields or variables are described for type (alphanumeric or numeric), label and size.

### *3.4.4 Database Validation*

After database design, any existing errors need to be identified and corrected. As the database design is a crucial step in conducting a study, the rule of “do and review” is always applied. The objective of this procedure is to find all possible errors among the database prior its transition to production environment.(9)

Database validation is important because a poor database design may adversely impact not only data entry but also data cleaning, extraction, listing and analysis.(9) At Eurotrials, database design validation is performed by a second person that did not participate in the database design, as it is not recommended to have the same person doing both tasks. When the database validation and design is performed by the same person the review can be biased.

It is common, when study data is already available, to test the database with real data. All errors found by the reviewer in the validation process are then corrected, being the database updated by the database designer.

In order to ensure that all errors found are duly corrected, a database functionality report is released with the findings detected during the database validation and the database update with the corrections performed.

### 3.5 Data capture and Data entry

Historically, data capture methods have been restricted by the available technology. However, with the considerable improvements on software and hardware, in the last years, technology factors are no longer considered restrictive.(14)

The term “data capture” refers to the storage of clinical data onto a database in a consistent, logical fashion so that it can be retrieved and searched. Ultimately, data receipt and entry is necessary for a study to successfully produce a clinical database of sufficient quality to support or refute study hypotheses. The content of the database should reflect the investigator’s observations collected at the clinical trial site, and this process of data collection should not obstruct investigator’s clinical practice.(14)

Whether a study is a paper-based or EDC, the functionality of the tools, the design of the study and the skill sets of staff should be carefully considered.

#### *3.5.1 Workflow*

The flow of data should follow a logical path, although specific processes and steps may vary between studies and organizations. Concerning the general workflow of paper-based data entry, the CRF pages should be tracked or logged, then entered, cleaned and subjected to rigorous audit/inspection or quality control. The workflow for EDC studies may vary according to the CDMS software used and study specifications.(8)

#### *3.5.2 Data receipt*

Data may be received through fax transmissions, regular mail, express delivery companies, hand delivery by monitors or transferred by other electronic means. All data sent from Monitoring staff must come with a transmittal form that describes all the information sent.(8)

When study data is received, the CDM staff will compare it with the corresponding transmittal form. If the study data is according to the description of the transmittal form, then CDM staff acknowledge the receipt and send a copy to the Monitoring staff. If the study data is not coherent with the transmittal form, the Monitoring staff is contacted to correct the information recorded in the form, either by sending the missing study data or sending a corrected transmittal form. This procedure may not be applied if the study data is sent directly by the sponsor without passing through monitoring staff.

#### *3.5.3 Data entry*

When data from a clinical study is captured on paper CRFs, this data must be transferred to a central database for final storage. The process of transferring it from paper to electronic storage is called *data entry*.(9)

## Curricular training Report in Clinical Data Management

At Eurotrials, data entry is performed manually by trained staff called *data entry operators* (DEO) who input data from a paper CRF onto a central database via pre-set data entry screens using conventional keyboards.(14)

Data entry processes should address data quality. Thus, accurately transcribing the data from the CRF to the database is essential. Common errors in transcribing are due to typographical errors or illegibility of the value as recorded on the CRF.(8, 9)

CDM staff use the following methods for studies using paper CRF:

- Double data entry with third party reconciliation of discrepancies
- Double data entry with second person resolving discrepancies
- Single data entry with extensive data checking
- Single data entry with no review

### Double Data Entry

In double data entry method, one DEO enters all the data in a first pass, and then an independent second DEO enters the data again in another instance of the database. In this technique the two entries are made and both are stored in two separate instances in the database. After both passes have been completed, the CDMS compares the two DEO's versions of the data in order to identify and highlight all differences between these two versions.(8, 9, 14) Further reconciliation between these two versions may be achieved by either two methods:

- A third person, usually a data manager, after all inconsistencies in data were identified or flagged (usually by automatic means presented in CDMS) decides whether there is a clear correct answer (one of the entries may have a typo, for example) and if it can be validated and transferred to database production area or if there is a discrepancy that must be registered because the data value is illegible or is in some other way unclear.(9, 14)
- The second DEO, usually a more experienced one, calls for a judgement to be made between the two conflicting entries, or flagging the inconsistencies for further investigation by CDM staff. This method is known as "*heads-up second data entry*".(9, 14)

### Single data entry

Single data entry is a valid option when there are strong supporting processes and technologies in place to identify possible typos or errors because of unclear data.(9) Single data entry methods divide in the following:(8, 9)

- Single data entry with no review – although not recommended, situations may occur where one person enters data and the data are not subsequently reviewed.
- Single data entry with review – one person enters the data and a second person reviews the data entered against the source data in order to address concerns of higher error rate.

### *3.5.4 Paper Case Report Forms considerations*

Although specific data entry processes are not mandated by regulatory bodies or suggested by ICH guidance documents, data entry processes should be adapted according to the quality level needed for each data field.(8)

Double data entering often gives an error rate of 0.1 to 0.2%, and it is typically used if keystroke errors occur or if random errors are likely to have a heavy impact in the analyses (8, 9). Therefore, double data entry has long been used without question as a reliable method of transcription. The rationale for applying double data entry depends if the increase in accuracy outweighs the expense and the associated time delay in entering the data twice onto the system.(14)

To maximize efficiencies, one approach that can be used is the single-entering all fields related to comments or entering long-text fields, since these fields are significantly more difficult to enter accurately and are often subjected to a later listing review by CDM staff, being the remaining fields entered with one of the methods for double data entering.(14)

The legibility of a CRF in paper based studies usually carries problems to DEOs, who are trained to reproduce the CRF page content onto the database, making no assumptions about the data that they are entering. Nevertheless, CDM systems have a feature that allows DEOs to flag any data considered doubtful or illegible allowing a later revision by experienced CDM staff, thus reducing input time by DEOs and avoid effort duplication.(14)

### *3.5.5 Data entry screen design*

Data entry screen design is an important factor that has a significant impact in the speed at which data can be entered and also in the error rate(8, 14). For paper-based studies, data entry screens should follow the pages of the CRF, and may even be designed to appear identical to the paper CRFs. Clearly label entry fields and ensuring that entry screens have sufficient space to enter and view expected data, are also strategies to consider when aiming data entry errors reduction.(8) Therefore, the greater similarity of the entry screen with the CRF page, the easier will be the entry of the data in the correct field by DEOs. Nevertheless, experienced DEOs often key very quickly, barely glancing at the data entry screen.(14)

Another consideration when designing data entry screens is format consistent analogous fields, such as dates, along the entry screens, which should be inserted in the same format specified in the CRF. For example using "DD-MMM-YYYY" in a single field or "DD","MMM","YYYY" in three different fields. The latter format is vastly used, because it allows the collection of partial dates, the month and year to be collected even if the day is unknown.(14)

In order to reduce errors by DEOs, the use of codelists in the data entry screen is very frequent as it restricts the input of data to a limited number of keyed responses. When fields are associated with

codelists, they can be programmed in a manner that if the value entered is different from those specified in the codelist, the DEO would be readily notified to rectify the value entered.(14)

### *3.5.6 EDC considerations*

Although sites are typically entering and cleaning data, CDM activities are needed to guarantee that data is being entered and processed properly. These data management actions can include training site staff on EDC system or measurement of site progress in data entry.(8)

The growing adaptation of EDC systems as primary system for data collection has a great impact in the training and desired skills for data entry staff. In a traditional data entry method, such as double data entry from paper CRFs, the skill emphasis on DEOs is the number of keystrokes made and the training emphasis is in the data entry system. Although, for EDC systems utilizing single entry, an overall understanding of the study becomes much more important in avoiding data entry errors. When EDC is the selected collection method, the data is entered directly onto the database by a member of the investigational team at the clinical site. Site staff have the capability to solve potential errors as data is entered, since the discrepancies are immediately identified as soon as they are entered.(8)

### *3.5.7 Modifying data*

Initial data entry is not the only data entry task performed by data management staff at Eurotrials. Posteriorly to the initial data entry, there are changes or corrections that may be performed. The processes needed for data correction are well-defined. Corrections are usually performed by CDM staff. Any changes after initial data entry, made by any person, are recorded in the CDMS audit trail.(9) The *Food and Drug Administration* (FDA) requires audit trails to record changes made in clinical data (21 CFR Part 11), and it should be possible, at any time, to check this audit trial.(15)

## **3.6 Electronic Data Capture**

Electronic Data Capture (EDC) systems deliver clinical trial data from the investigation sites to the sponsor through electronic means (computer programs) that replaces the paper CRF. The investigational site may enter the information directly into screens without first writing it down on a paper CRF. Nevertheless, the site may first record the information on paper and then enter it later. In both cases, the paper version of the data may be the normal site source documents or special worksheets provided by the sponsor. However this paper is not considered a CRF, and therefore it is not sent to the sponsor.(16)

Good quality EDC systems allow the data management process of the entire clinical trial. Therefore, several pharmaceutical and biotech companies have set goals in terms of the proportion of trials they wish to be

performed using EDC, moving from pilot studies to general adoption of this technology, as proof of confidence in EDC systems increase. (9, 17)

EDC systems are optimized for site activities during a clinical trial and typically feature:(9)

- eCRFs for the entry of data
- Tools for the sponsor, to raise automatic and manual queries for the discrepancies found while reviewing data
- Tools to allow sites to review and resolve discrepancies
- True electronic signatures so the investigator can sign-off the data entered in the EDC system
- Record or patient lock on the data
- Tools to assist monitoring
- Reports about patients for the sites and reports for the sponsor about sites
- A portal that provides information to the sites about the study
- A variety of ways to extract data for review and analysis.

### *3.6.1 Differences between Electronic Data Capture and Paper-Based studies*

Important areas that differ between EDC and paper-based studies are the manner in which data will be collected, the timeline necessary to prepare for the study and the manner in which collected data will be verified.

#### **Data capture methods**

At Eurotrials two approaches are used for collecting data:

- Online - The EDC method typically uses networked resources to record clinical data in electronic forms, which is only stored on a central server. There is no need to install software on the local machines of the clinical sites, and the investigators only need a computer with a browser installed and an internet connection. Also, Monitors, CDM staff and Clinical investigators have the possibility to access data from patients at any time and at the same time. This increases the level of control which is a great advantage when working in a heavily regulated environment.(8, 17)
- Offline – the traditional paper-based method for collecting and sending the data. (8,17)

#### **Study development and start-up timelines**

Many of the typical CDM start-up activities for both paper based and EDC studies include: Protocol approval, CRF design, CRF annotation, edit-check specification. The differences in CDM start-up activities for EDC studies are based largely on the increased number of tasks that must be completed before the study may begin.(8)

Several activities that have impact EDC study development and start-up timelines include(8):

- Role definition and access to data by authorized staff

## Curricular training Report in Clinical Data Management

- EDC system and trial specific training
- User account management
- Help desk support for users
- Preparation of coding dictionaries and processes as needed

### Source Document Verification (SDV)

The FDA has issued requirements for electronic records and signatures in 21 CFR Part 11, which provides criteria for considering electronic signatures as equivalent to handwritten signatures.(15) More important, conducting SDV on electronic records is the same as paper records. Before the start of an EDC study the database needs to be configured to support access, workflows and reporting requirements. Electronic records, like paper source records, must be accurate, original, legible, attributable and contemporaneous.(8)

### 3.6.3 Advantages and disadvantages of Electronic Data Capture

#### Main EDC Advantages

- Errors are caught in real-time. EDC alerts the clinical investigator that a suspicious data was entered, such as incorrect format or range or even when data is entered inadvertently. The investigator is informed immediately through a pop up that a suspicious entry must be verified. The final result is that the majority of the queries that would normally be encountered when using a paper CRF usually do not occur.(16)
- EDC systems, unlike paper, are able to get cleaner data up-front requiring less back-end cleansing, with the use of programmed heuristic edit checks leading to an highly reduction in the discrepancy management effort.(9, 16, 17)
- Reduction of the time from “*last patient last visit*” to database lock, as there is no need to perform a traditional audit to the database because there is no CRF to compare against the database searching for transcription errors.(16, 17)
- EDC enables investigator’s the submission of their own eCRF pages immediately after the patient’s visits, contributing to the general speed of the process and faster submission times. (16, 17)

#### Main EDC Disadvantages

Site management with EDC goes beyond the already challenging procedures for paper studies In addition to all the regulatory compliance requirements and protocol instruction, sites will need to have electronic signature procedures and forms, ongoing training in the EDC application, and be submitted to system accounts maintaining procedures.(9)

- Extensive preparation is required from Data Managers prior to the start of an electronic clinical study because it requires more upfront preparation than paper studies. Also, unlike paper CRFs, an



eCRF requires that the sponsor provide instruction to clinical investigators on how to use the computer program and how to get help if a problem occurs.(9, 16)

- EDC systems can be complicated and difficult to use by clinical investigators, hindering the advantages of EDC systems, so the best way to ensure is by taking a *test drive* with clinical investigators or get into consideration references from clinical investigators that already use these systems. Like any other technology, the easier the EDC system is to use, the greater the productivity will be across every part of the process.(17)
- Correction of bugs on the eCRF must be carefully planned and performed because it can affect the study data already collected, leading to possible data losses or down time on the eCRF for the sites.(17)

### 3.7 Data Validation Procedures

The aim of Clinical Data Management (CDM) is to guarantee timely delivery of high-quality data that are necessary to comply with good clinical practice (GCP) requirements, statistical analysis and reporting requirements. CDM validation activities play a pivotal position within drug development program and have a direct impact on the quality of data presented in a *New Drug Application (NDA)*.(18)

There are always discrepancies and data errors that are introduced into clinical database, no matter how careful clinical data is entered or collected. The majority of these data inconsistencies and errors can be greatly reduced or even totally eliminated by cautious review and data-validation activities.(8)

Data validation is defined as a number of steps needed to turn the original data collected in the CRF into a *clean/validated* database by checking discrepancies and resolving them. These steps should ensure that the database is accurate, consistent and a true representation of the patient's profile. (8, 9, 18)

#### 3.7.1 Defining and Implementing Edit Checks

Prior any data validation procedure, a Data Validation Plan (DVP) is developed by CDM staff. The DVP lists all checks to be implemented in order to detect discrepancies in study data.

Discrepancies are any inconsistencies in the clinical data that requires further clarification.(18)

Discrepancies can be as following:(9)

- Missing values (e.g., temperature values missing)
- Simple range checks (e.g., temperature value is not between 36°C and 38°C)
- Logical inconsistencies (e.g., no vital signs were collected but temperature has a value)
- Checks across modules (e.g., concomitant medication has given as solution to an adverse event, but no adverse event was recorded in the appropriate section)

## Curricular training Report in Clinical Data Management

- Protocol violations (e.g. vital signs were measured after the drug administration, and *per protocol* vital signs should be measured before drug administration)

In order to identify these discrepancies different methods are used by CDM staff: (9)

- Manual review of data and CRF's
- Computerized checks of data using CDM/EDC systems
- Computerized using external systems using SAS and EXCEL macros.

These checks used to find discrepancies are commonly named as "Edit checks". The main goal of edit checks is to draw attention to data that are inconsistent or potentially erroneous, which is triggered by data that are missing, out of range, unexpected, redundant and incompatible or discrepant with other study data.(8)

The edit check definition is specified in the DVP in order to implement the same checks across all clinical data during the course of study.(9,18) The DVP document usually assumes the form of a table with one row per check. Each check row is composed by the following columns:

- Module(s) where it is stored the data to be checked
- Type of check (manual or computerized)
- Check ID (unique ID code)
- Edit check description and message
- "Performed" (answer "Yes"/"No") to be filled by the assigned person: Programmer, if automatic check or data manager if manual check
- "Tested" (answer "Yes"/"No") to be filled by the assigned person that will review/approve the programming developed by the Programmer; usually this task is performed by the data manager or the quality controller

The CRF page number(s) and module(s) allow identifying where the value is being checked. To each check is attributed an identification code, allowing easy identification of the check. The logic operations of the edit check is written in a way that other collaborators can understand; usually symbols are used to abbreviate intensions (< lesser-than; > greater-than, = equal; <> different). It is also defined the message that will appear when the check finds a discrepancy. The 'type' column defines the nature of the check: manual or computerized. In each check specified in the DVP it is necessary to guarantee that it was performed by the appropriate CDM staff and it is correctly implemented.

### 3.7.2 Managing Discrepancies

The discrepancies found in edit checks are reviewed by a data manager familiar with the edit check specifications and with the CRF. Most of this discrepancies can be internally solved by CRF inspections and related data. The remaining discrepancies are sent to the clinical investigator site as *queries* for resolution.

## Curricular training Report in Clinical Data Management

Discrepancies that can be solved by data management staff are often called self-evident corrections (SECs). These corrections are restricted to a predefined list and limited to a specific study.(9) Also, these corrections should not imply any changes to actual data or results. Common examples include:

- Adverse events are recorded in the appropriate section, however the box asking “Are there any adverse events?” could be not filled. CDM staff can mark the box, however the inverse is not allowed
- Dates occurring in January are accidentally recorded in the preceding year. When data managers don’t have any doubt about the correct year, they can change it.

SECs can be managed either by informing the clinical investigators at the end of study about the changes that were made, or at the start of the study, when is provided to the clinical investigators a list with the allowed SECs.(9)

It is important to understand that SECs do not follow specific data entry guidelines. It is necessary to raise the discrepancy first. The change in the data is clearly identified and it would be necessary to give the reason for change, through this method the correction will be recorded in the audit trail associated with a person. (9)

### *3.7.3 Turning discrepancies into queries*

Any discrepancies that cannot be resolved internally by CDM staff, after thoroughly reviewed, are eligible to be sent to the investigator site for resolution. These discrepancies are usually sent in a specific form which has different names: “query form”, “discrepancy clarification form” or “data clarification form” (DCF). These forms have all the same purpose and they present the discrepancy to the investigator in an understandable and clear way in order to solve it.(9) A special care should be taken into account with the text that describes the discrepancy. Investigation site staff should easily identify the discrepancy that is under resolution. Also, when this text is being crafted it should be avoided leading questions. For example “Was 150 mg the administered dose?”, instead the message should be “Please clarify what dose was administered”, by this means the investigator staff needs to check the source document and not automatically answer ‘Yes’ or ‘No’.

### *3.7.4 Queries Resolution*

Once issued, the query form is sent to the Clinical Research Associate (CRA), via paper mail, or e-mail, being the CRA responsible to its delivery to the study sites. For paper based studies, the investigator staff typically answers into a spot on the DCF and posteriorly signs and dates it. DCF should be treated as CRF’s as they contain original site data. When a site provides the resolution to a discrepancy on a DCF, they may write a paragraph explaining the correct action or justifying the value, and data management staff needs to

interpret it in order to identify what action should be taken. Sometimes it is necessary to recheck or resend the DCF to the investigator site to clarify the resolution provided, due to problems in interpreting the response, or because the resolution still carry some degree of inconsistency.(9)

Those queries that require a change on data or provide a missing value are applied by a different path from the initial data entry, the same path as SECs.(9)

Not all queries result in data changes; the data may be as it was reported in the CRF, or were not collected at the time and nevertheless the responses must be recorded in the discrepancy management system to close the discrepancy.(9)

After changes on data were made during queries resolution process, it is essential to rerun all validation rules over the data, as it is very common for data updates to cause some other discrepancy, as part of discrepancy resolution.(9)

### 3.8 Database Quality Control

Proof of data quality is essential for meeting regulatory requirements. Thus, data collected during clinical trial must have as few errors as possible to support findings or conclusions drawn from that trial.(8)

At Eurotrials, CDM staff performs database audits to fulfil the Quality Control (QC) procedures planned. Commonly for a specific study, the quality control staff do not participate directly in data management activities. The first step to assemble a QC plan is to identify the CRFs that should be used (data sample definition). Then these CRF pages are collected as well the associated query forms. After that, the data is compared with the one stored at the central database. The audit result is usually given as the number of errors against the number of fields on the CRF or database (calculation of error rate). Depending on the result achieved in relation with the maximum error rate defined in the QC plan, corrective actions may be taken (8). Finally, the QC procedures should be documented.

#### 3.8.2 Data Error Categorization

Database quality control relies on the assumption that errors found in the database can be categorized in the following groups:(8)

- *Error type A:* error not leading to misinterpretation of the meaning/value on the study data information
- *Error type B:* error leading to misinterpretation of the meaning/value on the study data information
- *Error type C:* any error type B found on study objective variables

### 3.8.2 Data sample definition

The most frequently used number for database quality control is to audit all the data of 10% of the patients that completed the study. According to what was defined in the QC plan, this number can be supplemented by a 100% audit of all patients' safety fields (such as those for Adverse Events) and/or a 100% audit of all patients' key efficacy fields.(8)

### 3.8.4 Error rate calculation

Calculating an error rate is a better approach than counting the number of absolute errors because it facilitates the comparison of data quality across database tables and trials.

The error rate is defined as the number of errors detected divided by the total of fields inspected.

The following table represents the acceptable error rates:

**Table 4 - Acceptable error rates**

	Type A	Type B	Type C
Single data entry	Not applicable	$\leq 0,05$	$\leq 0,02$
Double data entry	$\leq 0,10$	$\leq 0,01$	$= 0,00$

### 3.8.4 Corrective actions

Any errors found in the database during the quality control are corrected by quality control staff.

If the error rate of the selected sample does not exceed any of the acceptable error rates, then the database quality control is concluded. Otherwise, if at least one error rate is exceeded, a new sample of data is selected and the procedure is repeated until the errors found in the selected sample do not exceed any of the acceptable rates. In extreme cases, this procedure could lead to an audit to all study data (100%) of all the patients that concluded the study.

### 3.8.5 Documentation

After the conclusion of database quality control, the study data manager and the data manager quality controller will release a database quality control report detailing all errors found in the comparison between the study data and the database. Anyone that reads this report should be able to recreate the sampling and error rate calculations and produce the exact same results.

### 3.9 Final Database

After the last patient's data has been collected from the sites and all data is validated, the study database is ready to be closed and locked in order to ensure data integrity. Once the database has been locked, the study database is considered *final*. (8, 9) The last responsibility of CDM staff is to deliver the final database to the statistics staff, so that the final statistical analysis can be performed and study conclusions drawn.(8)

#### 3.9.1 Database Closure Procedures

Closing and locking a study database is of vital importance as it prevents inadvertent or unauthorized changes. Although important in open label studies, database closure and lock plays a role even more important to preserve the integrity of randomized trials after the blind has been broken. Thus, there is a well-defined process for closing the study database, which is followed to ensure that relevant study staff have been informed or approved the database lock.(8)

The preparation for database closure includes the following procedures:

- All data have been received, processed and validated
- All queries have been resolved
- Reconciliation of sponsor adverse event database, if existent, with the main study database
- Conclusion of database quality control, with acceptable database error rates and documentation of the errors that occurred
- All documentation is updated and stored according to standard operation procedures.

After the conclusion of database closures procedures, the database will be locked by CDM staff and considered as the "Final Database". Then, a signed certificate by the CDM staff is released (and sent to sponsor, if required). As soon as this certificate is signed, edit access to the database is removed to all other users by the study data manager.

#### 3.9.2 Errors found after database closure

After database lock, if an error is found, the sponsor is readily notified. CDM staff along with a statistician and/or a medical reviewer evaluates the potential effect of this error on the analysis of safety and efficacy.

Not all errors found must be corrected in the database itself, being these errors just documented in the statistical or clinical report. It is the sponsor who decides to change all errors found in the study database or only to change those that have major impact on the safety/efficacy analysis.(8)

If it is decided to eliminate any error, a listing containing the subject identification, item identification, value to be replaced, the new value and the reason for data replacement, will be provided to the CDM staff. The

changes performed to the database after the un-lock are documented (and sent to the sponsor, if required). Re-locking the database follows the same process for notification as the initial lock.

### *3.9.3 Randomization code break*

In case of a blind-study, the randomization codes must only be broken after the study database lock. The only exception is in the case of a medical emergency. The un-blinding of the randomization codes must be authorized by the sponsor.

The randomization codes are entered by DEO in a specific dataset which is reviewed by CDM staff. This dataset is essential to assign the treatment arm with the investigational medical products and comparators so that a correct safety/efficacy analysis can be performed.

## 4. Data Standards and CDISC Implementation

### 4.1 Data Management Standards in Clinical Research

The advent of modern information technology has enabled widespread use of comprehensive standards that encompass almost every part of data collection and handling.(8) Although there are few regulatory mandates for using any particular standard, using standards in all areas of data collection and handling can greatly increase an organization's efficiency by shortening reducing overall time and expenses while maintaining consistency for data managers and those charged with collecting data at clinical sites.(8,19) Most of the established standards currently in use are readily available and designed to be independent of any vendor or platform.(20)

Although multiple standards exist for similar concepts, the ultimate goal is for researchers everywhere to use the same standards and naming conventions for their studies. This goal has not yet been reached, but the clinical research industry is trending in that direction. If different parts of an organization can freely use whatever data standards as they wish, many of the benefits of standards are lost.(8)

#### *4.1.1 Barriers to Standards Implementation*

There are a lot of wrong assumptions taken by corporations about standards implementation that are nothing more than myths, consider the following:(19)

- Standards stifle creativity
- Companies have their own way of doing things so standards don't apply to them
- Standards are always changing so companies should wait until they stabilize
- There are too many standards so it is too hard to choose
- Standards can be implemented at the last minute when required without any change to the existing organization
- There is no good time to implement standards

#### *4.1.2 Purpose and Benefits of Standardization*

Within the context of CDM, standards are used to optimize the collection, transport and storage of data, and to simplify the submission of data to regulatory bodies.(8)

The use of standards within clinical research involves using standardized names, coded, structures and formats for data across different locations, studies and organizations. Using the same formats, names and codes for different studies can greatly decrease the time and money needed to set up a study, particularly in cases where similar studies have been conducted in the past. (20) Standards provide benefits beyond



study setup and can also help streamline processes for study conduct, data transfers, analyses and regulatory submissions.(19, 20) Ultimately, standards facilitate bringing safe and effective treatments to patients in a more timely and cost-effective fashion.(19, 20)

It is clear that data standard bring more than just savings and a reduced cycle time, but companies may also see the following benefits to their organization when they implement standards:(19)

- Communication among project teams and partners becomes easier, regardless of which part of the study they participate.
- People working within the clinical research environment execute the work with more accuracy and less up-front training as the process is consistent and not constantly changing.
- Decision making becomes simpler, as format or data fields are already pre-specified.
- Enables the use of different technology/tools as long they are compliant with the same standards
- Easier transfer of data between partners.

### *4.1.3 Data Standards categories*

Data standards can be divided in the following categories:

**Standards Governance** – if every part of an organization were free to use any standards they wish, many of the benefits are lost. This would end with the proliferation of redundant standards and each part of the organization would be forced to develop their own databases and analysis tools. In order to avoid this, an organization needs a standards governance structure that specifies what standards will be used by the organization. As much as possible the governance board should adopt standards that are widely recognized by a broad standards community (e.g. ISO, ANSI).(20)

**Exchange Standards** – refer to standards for structuring data for exchange between organizations. An exchange standard describes how to organize the data for exchange between organizations. An exchange standard does not specify what data should be exchanged (content) neither terminology used to describe the content.(20)

**Terminology Standards** – a critical part of any implementation of a data standard is the terminology used to describe the data. While data attributes are often described by a single word, a rigorous terminology standard requires a concept definition, a code for the concept and a list of labels associated with the concept. While different users use different labels for the concept, the concept definition and concept code remain unchanged.(20)

**Content standards** – are standards that define what information needs to be exchanged between organizations. An organization may need a large amount of diverse content, but a particular user may need only a fraction of the data that could be potentially exchanged via the standard.(20)

**Reporting and analysis standards** – a variety of datasets are needed by different users of the data. Some users will need specific static reports, while others will need general purpose datasets that can be used for analysis either pre-specified or *ad-hoc*. Unlike exchange standards, reporting and analysis standards need to be tailored to the specific use planned by the end users.(20)

**Data acquisition standards** – ideally data should be collected in a way consistent with the way it will be exchanged. Data should only need to be entered once, recoding should be kept to a minimum and every piece of data should be mapped to a location in the implemented standard.(20)

### 4.2 Clinical Data Interchange Standards Consortium Background

Clinical Data Interchange Standards Consortium (CDISC) and was formed solely to create standards for clinical/medical research data. In their mission, states “*CDISC is a global, open, multidisciplinary, non-profit organization that has established standards to support the acquisition, exchange, submission and archive of clinical research data and metadata. The CDISC mission is to develop and support global, platform-independent data standards that enable information system interoperability to improve medical research and related areas of healthcare. CDISC standards are vendor-neutral, platform-independent and freely available via the CDISC website*”.(21)

CDISC was created in 1997 to address the needs of FDA and biopharmaceutical needs. However, today CDISC is making a pronounced effort to ensure that the developed standards are broadly applicable to medical research whether it is academic, government-sponsored or for product development.(8,19)

It is important to refer that since the inception of CDISC, all standards developed were performed by volunteers.(8,19)

CDISC works closely with other standard development organizations such as ISO and HL7. This collaborations demonstrate the commitment to improve interoperability among clinical and research systems as well as a closer relationship between clinical and research standards.(8, 19)

CDISC in order to streamline medical research from the protocol through reporting of results while improving data quality and patient safety created many standards such as:(8,19)

- Clinical Data Acquisition Standards Harmonization (CDASH) – Standard that defines a minimum set of data collection fields in CRFs
- Study Data Tabulation Model (SDTM) – The content standard for regulatory submission of case report form tabulations from clinical research studies.
- Case Report Tabulation Data Definition Specification – Also known as “define.xml”, defines the format and content of specification and attributes of CDISC SDTM datasets.
- Analysis Data Model (Adam) – The content standard for regulatory submission of analysis dataset and associated files.
- Terminology – The controlled standard vocabulary and code sets for the all CDISC model/standards

### 4.3 Study Data Tabulation Model Overview

The Study Data Tabulation Model (SDTM) provides a standard structure for describing information collected during clinical trials that are to be submitted as part of a product application to the regulatory authorities.(22, 23) The SDTM is used either to submit clinical trial or non-clinical data for product application across all therapeutic areas.(22 ,23) The SDTM model has two important ways to provide flexibility. First, for each domain/dataset it is defined a superset of variables (an exclusive list) where only a small number of variables are required to be present in every submission. Second, the set of domains/datasets required to a specific application is not defined by the model but determined with the goals and endpoints of the clinical trial protocol.(22, 23)

The availability of defined standard data for submission provides many benefits to the reviewers as they can be trained in the principles of standardized datasets and the use of standard software tools, and thus be able to work with data effectively with less preparation time.(22, 23)

The SDTM is based on a concept of observations about subjects who participated in a clinical study. An observation is described by a series of variables corresponding to a row in a dataset or table.

Most of the subject-level observations collected during the study are represented in one of the three SDTM general observation classes:(22, 23)

- The Interventions class – captures investigational, therapeutic and other substances that are administered to the subject either as specified by the protocol (e.g. exposure to study drug),

## Curricular training Report in Clinical Data Management

coincident with the study assessment period (e.g. concomitant medications), or self-administered by the subject (e.g. use of alcohol, tobacco or caffeine)

- The events class – captures planned milestones such as randomization and study completion, and occurrences, conditions, or incidents independent of planned study evaluations occurring during the trial (e.g. adverse events) or prior to the trial (e.g. medical history).
- The Findings class – captures the observations resulting from planned evaluations to address specific tests or questions such as laboratory tests, ECG testing, and questions listed on questionnaires.

Observations about study projects are usually collected for all subjects in a series of domains. A *domain* is defined as a collection of logically related observation with a common topic. In non-standardized studies the domain is known as *modules*. Each domain is represented by a single dataset and is distinguished by a unique (except for Relationship Datasets) two character code that is used consistently throughout the SDTM.(22, 23)

There are standard domains and their respective codes that are already defined by CDISC team which falls in the following seven categories:

**Table 5 - SDTM Domains(22, 23)**

Special-Purpose Domains	
Demographics - DM	Comments - CO
Subject Elements - SE	Subject Visits – SV
Interventions General Observation Class	
Concomitant Medications - CM	Exposure - EX
Substance Use - SU	
Events General Observation Class	
Adverse Events - AE	Disposition - DS
Medical History - MH	Protocol Deviations - DV
Clinical Events - CE	
Findings General Observation Class	
ECG Test Results - EG	Inclusion/Exclusion Criterion Not Met - IE
Laboratory Test Results - LB	Physical Examination - PE
Questionnaires - QS	Subject Characteristics - SC
Vital Signs - VS	DRUG Accountability - DA
Microbiology Specimen - MB	Microbiology Susceptibility Test – MS
PK Concentration - PC	PK Parameters - PP
Findings About	
Findings About - FA	

## Curricular training Report in Clinical Data Management

Trial Design Model	
Trial Arms - TA	Trial Elements - TE
Trial Visits - TV	Trial Inclusion/Exclusion Criteria - TI
Trial Summary - TS	
Relationship Datasets	
Supplemental Qualifiers – SUPPQUAL or multiple SUPP-- Datasets	Related Records - RELREC

When preparing the SDTM for a specific study, only domain datasets that were collected or even derived data from the collected data are submitted. The collected data for a specific study can use some or all the standard domains developed by CDISC team as well additional custom domains based in three general observation classes.(23)

All SDTM variables allowed for each domain are defined. Generally, for each domain the variable is named with its two letters code prefix followed by a word fragment, allowing easy identification of the variable origin.(22, 23) The SDTM model does not foresees the addition of custom variables on the existing domains. These variables are recorded as values in the *SUPP Domain*. Furthermore, it is not necessary to represent all variables created by CDISC team when submitting SDTM to regulatory authorities. Thus, CDISC team created the “Core” concept to aid in variable compliance. The three categories specified are:(23)

- Required – these variables must always be included in the dataset and cannot be null for any record. Required variables are necessary for identification of a data record or to make a record meaningful.
- Expected- these variables are necessary to make a record useful in the context of a specific domain. When no data were collected for an expected value, a null column should still be included in the dataset, with a comment in the metadata (“define.xml”) stating that data was no collected.
- Permissible – these variables are used in a domain, as appropriate, when collected or derived. When data is not collected for these variables, there is discretion to either include them or not, as a column. However, there is no discretion to not submit permissible variables when they contain data.

Each variable usually takes place as a column in the dataset and following the SDTM Model is classified in the following five major roles:(23)

- Identifier variables – identify the study, the subject, the domain and the sequence number of the record
- Topic variables – specify the focus of the observation which vary according to the type of observation (laboratory test, medication treatment, or medical history term)

## Curricular training Report in Clinical Data Management

- Timing variables – describe the timing of an observation (start date, end date or time of collection)
- Qualifier variables – include additional text or numeric values that describe the results or additional characteristics of the observation (such as units or descriptive adjectives)
- Rule variables – expresses an algorithm or an executable method to define start, end or looping conditions only applicable in the Trial Design model.

Within a domain, CDISC team defined the order of the variables, in order to facilitate the review. Variables for the three general observation classes should be ordered with *Identifiers* first, followed by the *Topic*, *Qualifier* and *Timing* variables.(23)

If the following observation example is taken: “Subject 01 had 35 °C of temperature starting on day 5”, the value for *identifier* variable is “Subject 01”, the *topic* variable is “fever”, the *qualifier* variable “35 °C” and the *timing* variable is “starting on day 5”.

Dataset or domains prepared for submission are described by metadata definitions that provides information about the variables used for each dataset. The metadata is described on DDD usually named “define.xml” and it is submitted with the datasets to regulatory activities.(24)

The “define.xml” has two main parts: Table of Contents (TOC) and Data Definition Tables (DDT). The TOC lists all datasets/domains included in the submission and the DDT describes the variable level attributes, definitions and usage information for each variable contained within each domain.(24)

The Table of Contents contains the following fields to describe the metadata of SDTM domains/datasets:

**Table 6 - Define.xml Table of Contents(24)**

Field	Description
Dataset	The name of the domain/dataset (e.g., “AE”).
Description	A short description of the type of information contained within the domain/dataset (e.g., “Adverse Events”).
Structure	The level of detail represented by individual records in the dataset (e.g., “One record per subject”, “One record per subject per visit”)
Purpose	Purpose for the dataset (e.g. tabulation)
Keys	Key variables are used to uniquely identify and index record in a domain/dataset.
Location	Folder and filename where the domain/dataset can be found.

## Curricular training Report in Clinical Data Management

The Data Definition Table section contains the following fields to describe the metadata of SDTM domains/datasets:

**Table 7 - Define.xml Data Definition Table(24)**

Field	Description
Variable	The field name of the variable.
Label	A brief description of the variable.
Type	The variable type (Character or Numeric).
Controlled Terms or Format	The set of controlled terms or variable display information.
Origin	Indicator of the origin of the variable (e.g., CRF pages, derived, protocol).
Role	Information on how a variable is used within the dataset (e.g., "Identifier", "Topic", "Timing", "Qualifier", "Rule").
Comment	Other information regarding the variable definition, usage.

FDA has referenced the use of the SDTM in the Study Data Specifications and "define.xml" for the Electronic Common Technical Document. In spite of the use of these standards are recommended, these models have not been mandated. Additionally, Study Data Specifications provide other useful technical instructions for submitting clinical and non-clinical data, as CDISC standards strictly complies with file format and variable/data size.(25)

The transport file format used in product application for FDA is the "SAS XPORT", known as "Version 5 SAS", which is an open format. Thus, data in this format can be converted to and from this SAS transport format to other commonly used formats without the use of programs from SAS Institute or other specific vendors.

The maximum size of variables and data is as follows:(25)

**Table 8 - Maximum size of variables and data(25)**

Element	Maximum Length in Characters
Variable Name	8
Variable Descriptive Label	40
Dataset Label	40
Data	200*

\*However for all datasets the maximum number of characters is 200 for data values, the allocated character number should be the maximum length used in the values collected. For example, if the Subject Identification field has a maximum of 18 characters, then, when building the database, the length should not be set to 200 but to 18 for that field. This procedure allows reducing dataset file sizes.

### 4.4 CDISC Standards Implementation

Eurotrials' Clinical Data Management Department is now piloting the implementation of SDTM, "define.xml" and CDISC terminology in completed legacy clinical studies *phase I* and *phase II* of the same investigational product. Legacy studies are studies that were conducted without taking into consideration CDISC standards.

However, the Eurotrials' implementation of these standards started with a late legacy studies conversion, being the ultimate objective, to implement CDISC standards in new studies earlier in clinical development of investigational products in order to usufruct from the benefits of standards usage.

Eurotrials' implementation of CDISC standards on legacy data is performed in four phases:

#### Phase 1 – Process and Data Review

All databases from the different clinical trials, as well study protocols, annotated CRFs and data management plans were collected and compiled with the objectives of creating an inventory of the documentation and its posterior review. The goal of this review is to analyze the gap between non-normalized datasets against the CDISC SDTM model and identify the disparate definitions.

#### Phase 2 – Detailed Analysis and Design

After the revision of the compiled documentation, it is defined for each clinical study the necessary data transformations and migration to normalize the existing datasets. The output of this definition is mapping the non-normalized datasets to a repository where all datasets and variables needed comply with SDTM rules.

A plan to load the legacy data on the new SDTM datasets following the proposed mapping will then be performed, this plan will also include the intended tools and software to be used as well any validation required regarding the software.

This phase is the most consuming one and is of vital importance to the success of CDISC standards implementation.

#### Phase 3 – Implement Tools

It is performed in a small scale the transformation and migration plan using the defined and validated tools. The converted datasets are thoroughly reviewed in order to assess if the plan was successfully implemented.

Before proceeding to the next phase, all steps and procedures need to be fully validated and all difficulties or bottlenecks identified.



## Curricular training Report in Clinical Data Management

### Phase 4 – Production

All data on legacy studies are transformed into SDTM datasets and the process documented. Eurotrials will perform a final report and prepare documentation for posterior traceability. The data is then delivered to the sponsor in the standard form using SAS transport files complemented with the “define.xml”.

## 5. Curricular Training Timeline

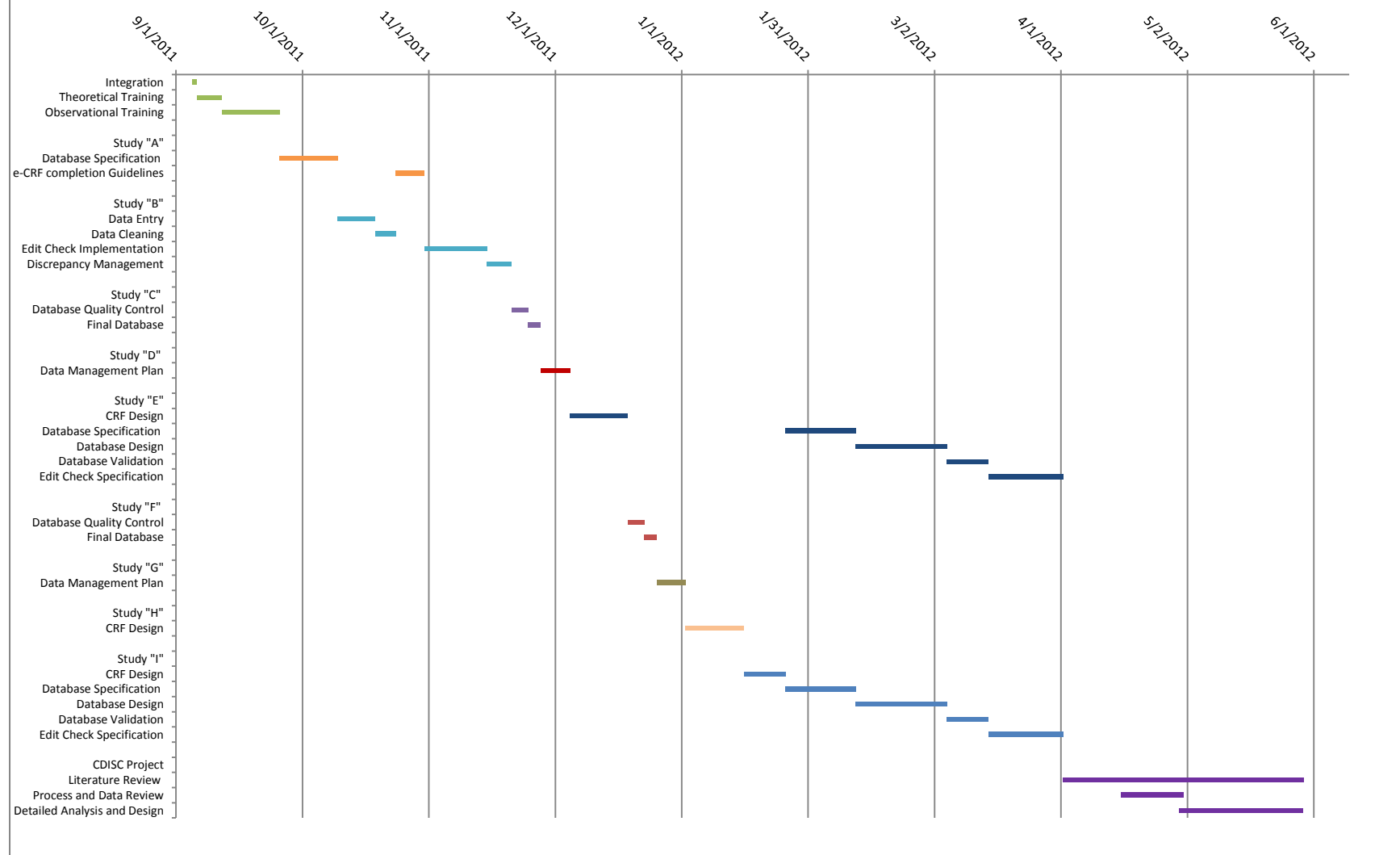


Figure 5 - Curricular Training Timeline

## Curricular training Report in Clinical Data Management

The curricular training in Clinical Data Management started in September of 2011. The occupied position to perform the curricular training in the Data Management Department was the Database Operator which is in terms of responsibility and autonomy between the Data Manager and Data Entry Operators. Despite of Database Operators having fewer activities that can be performed autonomously compared to Data Managers, Database Operators can perform activities usually assigned to Data Managers, if these performed activities are then reviewed by Data Managers to assure that they were properly executed by Database Operators.

The integration of a new collaborator at Eurotrials is performed by the Human Resource Department Staff. The integration procedures begin with a training in Eurotrials basic procedures and the reading of the "New Collaborator Guide".

Since Eurotrials deals with confidential documents and data, as well proprietary procedures the new collaborator signs a confidential agreement to prevent the disclosure of confidential and proprietary information to third parties

It is given to each collaborator a personal file, that includes its Curriculum Vitae and it is registered all the training performed whether its job or project specific.

These procedures are then followed by a formal introduction of the new collaborator to the current collaborators from all Eurotrials departments.

The training of a Database Operator starts with a job specific training in the theoretical basis of clinical trials and CDM activities. Eurotrials provides the bibliographic material which is composed by the Eurotrials Standard Operation Procedures and a set of literature, including books, guidelines, software manuals and tutorials.

Followed by the job specific training every time Clinical Data Management staff is assigned to a new project the corresponding training activities are implemented, including training on the therapeutic area of the study, training on the project study protocol and CRF.

Any training activities are recorded in the appropriate form and archived in the collaborator's personal file.

Following the initial theoretical training, an observational period of the activities performed by the CDM took place in order to understand the flow of clinical data and interaction between the different departments.

## Curricular training Report in Clinical Data Management

During the 9 months training period I actively participated in the following studies:

**Table 9 - Studies Description**

Study	Description
Study "A"	Ovarian Cancer – Phase II, EDC, local, multicenter and open label
Study "B"	Ovarian Cancer – Phase II, paper-based, local, multicenter and open label
Study "C"	Osteoarthritis – Observational Study , paper-based and local
Study "D"	Ophthalmology – Phase IV, EDC, multinational and open label
Study "E"	Prostate Cancer – Phase II, EDC, local, multicenter and randomized
Study "F"	Persistent Organic Pollutants in Breast Milk - Observational Study , paper-based and local
Study "G"	Purple Immune Thrombocytopenia - Observational Study, paper-based and local
Study "H"	Growth Hormone Disease in Children – Phase III, local, multicenter and randomized
Study "I"	Prostate Cancer – Phase IV, EDC, local, multicenter and randomized
Study "J"	Implementation of SDTM, "define.xml" and CDISC terminology on legacy studies.

One of the first activities executed during the curricular training was the specification of a database for the Study "A". To perform this activity it was necessary to understand the e-CRF structure and design, as well EDC platform basic functions. Firstly it was performed the DDD which contained a list of all pages, modules, variables and associated codelists as well the description of the attributes of these items. To conclude the specification it was performed a annotation of the e-CRF was done, since this oncology study used the EDC platform it was necessary to extract the e-CRF pages to a paper CRF so that the annotation of the pages, modules, variables and codelists could be executed. The specification of database is frequently required by the statistician to develop the Statistical analysis plan. This activity was also performed in the "E" and "I" studies

In Study "A" it was necessary to create an e-CRF completion guideline that could be provided to the clinical investigators and monitors. To perform this task it was necessary a serious training to the EDC platform to ensure that the necessary features were clearly explained. The key topics explained in this document were the process of obtaining a valid user and password, the log-in process, creation and management of patients, data capture and discrepancy management.

It was on Study "B" that I had the first contact with CDMS. I performed DEO activities for this study and introduced CRF laboratory data in the database via CDMS. Since in this study was paper-based all data were introduced by double data entry method. As Database Operator I cleaned all discrepancies found with the reconciliation of the two DEO databases.

## Curricular training Report in Clinical Data Management

I have also participated in edit checks implementation in order to validate the study “B”. The majority of the automatic edit checks were programmed by the SAS programmer. However, the remaining edit checks that were not programmed using SAS were implemented by me using the Excel tool. Further to the automatic implementation of the edit-checks I have also implemented manual checks defined at the DVP which required to accurately review the paper CRF.

Any discrepancy found either by automatic or manual check was then reviewed by Data Manager. If the Data Manager considered that the discrepancy could not be solved internally, then the discrepancies were transformed into queries to be sent to the Eurotrials Clinical Monitors. I have participated in this process of sending DCF’s to the investigator’ site, as well I prepared the necessary documentation for DCF’s tracking such as transmittal forms.

For studies “C” and “F” I participated in the final Database Quality Control Procedures, checking the values in the database with the values present on paper CRF’s and DCF’s. The values to be verified were specified on the QC plan.

I cooperated with Data Managers responsible for the studies “D” and “G” to create the DMP, the main task performed by me was to extract the necessary information from the studies protocols, as well identification of assigned data management staff.

I have participated in the development of a paper CRF draft for an EDC studies such as “E”, “H” and “I”. The purpose of designing a paper CRF for a study intended to be implemented on an EDC platform is to get the approval of the Sponsor on which data is to be collected, how is collected and when is collected. When designing a CRF one of the most important steps is to carefully review the protocol as it specifies the primary and secondary endpoints of the study. The activity of designing a CRF is greatly supported by the medical writer.

I had the possibility to participate from the beginning in studies “E” and “I”, these studies were from the same sponsor, investigating the same investigational medicine product. However from distinct phases, there were similarities between them. Thus, excluding CRF design, the database building, database validation and Edit Check specification procedures were performed in parallel in order to take advantage from synergies.

My enrollment in these studies allowed me to acquire essential skills to any professional in the field of CDM. I have collaborated in the database construction for both studies, since the databases were built using the CDMS my knowledge in this software greatly increased. The built database reflected the CRF draft approved by the sponsor. As said previously I also performed the database definition for these studies.

Other important task learned with these studies was how to build a DVP and how to specify an edit-check. To perform this activity it is of extreme importance a good knowledge in the study protocol and CRF, as it is

## Curricular training Report in Clinical Data Management

the only way to cross relevant data from the CRF and guarantee that the procedures planned on the study protocol were followed.

My last task in both these studies was to validate the built databases, this task only occurred when the programmer concluded the set-up of the EDC system. Through the EDC platform it was verified if the behavior of the variables was the behavior that was previously specified. The final aim of database validation is to transfer the study set-up from test environment to production environment.

In April, I was integrated in the Eurotrial's team responsible to implement the CDISC standards. Prior to go hands-on the job, it was necessary to understand the concept of the existing CDISC models. The available CDISC literature is extensive and additionally to the official literature there are conferences materials that provide further enlightenment with real cases of CDISC implementation.

At the time of curricular training conclusion, the step 1 was completed and the step 2 was ongoing. Although the CDISC implementation steps were sequential, the task of reviewing the literature will go along with the all project since the official literature defines the intended outputs of the standards implementation

### 6. Conclusion

The 9-month curricular internship at Eurotrials in the Clinical Data Management Unit was a great experience that allowed me to learn and develop myself as a future data manager. This CRO has a great work environment and it's a place where all the collaborators and trainees have the chance to stand out and contribute to the development of Clinical research in Portugal. During my internship, I performed several data management activities that allowed me to identify this area, as the one to develop my career. I think that Eurotrials is a great place to develop this career and develop the required.

It allowed me to develop my team working skills, by interacting with several teams and areas of expertise. This constant interaction and exchange of information between all the departments allows the support and understanding of each activity's flow that is performed in the context of Data Management. For instance, in order to correctly manage the information that arises from a clinical trial several inputs are needed from the Clinical Trials Unit. After the treatment of this data, the information is sent to Biostatistics Unit, which requests certain types of information in a certain format. This work flow wouldn't be possible without the constant understanding and communication between all units.

Among other things, one of the best ones that I realized during this internship was the preparedness that the Master in Pharmaceutical Medicine and the Degree in Biomedical Sciences gave me. Many topics were reviewed and several were applied during the internship.

During the internship, I gained awareness of the pivotal role that CDM activities have in the drug development processes. It was interesting to verify the diminishing number of clinical trials performed in Portugal and how our working volume is shifting from Europe to central and Latin America.

The work environment and the personnel is one of the greatest assets of the data management unit. The persons that are part of this team are the ones responsible for the good working environment which largely helps the learning process and development of both social and professional skills. I am also grateful for the professional opportunities that Eurotrials are currently giving me, there is nothing more important than seeing our skills and competencies recognized.

After the internship I intend to continue to extend my skills and capabilities in clinical data management and proceed with the career progression planned by Eurotrials.

An obvious hurdle with the internship was the ability to conciliate the professional activities with the curricular ones; however I think that I was successful at managing the activities that I needed to do. Nevertheless the major hurdle was the adaptation to a new reality, because most of the times it can be very

intimidating and may be a step back, however this was overcome with the help of my friends and co-workers at Eurotrials.

## 7. References

1. Eurotrials - Scientific Consultants. 2008 [05/Dec/2012]; Available from: <http://www.eurotrials.com/>
2. Eurotrials. Quality Manual. 2012.
3. Krishnankutty B, Bellary S, Kumar N, Moodahadu L. Data management in clinical research: An overview. Indian Journal of Pharmacology. 2012. 44(2): 168–172.
4. Warrington S. Purpose and Design of Clinical Trials. The Textbook in Pharmaceutical Medicine
5. European Medicines Agency. Note for Guidance on General Considerations for Clinical Trials (CPMP/ICH/291/95). 1998.
6. Woodcock J, Woosley R. The FDA Critical Path Initiative and Its Influence on New Drug Development. The Annual Review of Medicine. 2008. 59:1-12.
7. The CRO Market Outlook: Emerging Markets, Leading Players and Future Trends. Scripps Business Insights. 2009.
8. Society for Clinical Data Management. Good Clinical Data Management Practices. 2011.
9. Prokscha S. Practical Guide to Clinical Data Management. 2007.
10. Avey M. Case Report Form Design. In: Clinical Data Management. 2000.
11. International Conference on Harmonisation. E6(R1) - Guideline for Good Clinical Practice. 1996.
12. Spilker B. Guide to Clinical Trials. 1991.
13. Wright P, Haybittle J. Design of forms for clinical trials. BMJ. 1979. 2(6190): 590–592.
14. Waterfield E. Data Capture. In: Clinical Data Management. 2000.
15. Food and Drug Administration. Guidance for Industry - Part 11, Electronic Records; Electronic Signatures - Scope and Application. (2003).
16. Collins S. EDC and the Changing Role of the Clinical Data Manager. Data Basics. 2007.
17. MacGarvey A. A snapshot view of Electronic Data Capture (EDC) - how far it's come, how far it has to go and how well it's been accepted. Innovations in Pharmaceutical Technology. 116-118.
18. Patel P. Data Validation. In: Clinical Data Management. 2000.
19. Montjoie AJ. Introducing the CDISC Standards - New Efficiencies for Medical Research: CDISC. 2009.
20. Levine J. Data Standards For Clinical Trials. Data Basics. 2011.
21. Clinical Data Interchange Standards Consortium. 2012 [05/Dec/2012]; Available from: <http://www.cdisc.org/mission-and-principles>.
22. Clinical Data Interchange Standards Consortium. Study Data Tabulation Model. 2008.
23. Clinical Data Interchange Standards Consortium. Study Data Tabulation Model Implementation Guide: Human Clinical Trials. 2008.
24. Clinical Data Interchange Standards Consortium. Case Report Tabulation Data Definition Specification (define.xml). 2005.
25. Food and Drug Administration. Study Data Specifications. 2012.